Project 2 Milestone

Haoen Huang, Liuyi Cui, Yujia Hong, Xiaoyuan Liu

March 2023

1 Abstract

In the genomics era, the availability of large biomedical datasets with genome-wide readouts has the potential to identity cancer vulnerabilities, differential cancer gene, and and classify predictive models. Integrating the datasets resulting from CRISPR KO study can help us assemble a comprehensive map of cancer genetic vulnerabilities. Here, we integrated the public independent CRISPR KO screens performed to Public 22Q4 Primary Files at Depmap website by machine learning algorithms. Our integrated datasets recapitulate findings from the individual datasets, provide greater statistical power to cancer- and subtype-specific analyses, unveil additional biomarkers of gene dependency, and test if the classification schemes derived by associating gene dependencies with other data sets such as RNA expression, pathway signatures, DNA mutation and copy number data are unique.

2 Introduction

Cancer is a complex disease that can arise from multiple different genetic alterations. In order to identify and prioritize new potential therapeutic targets for precision cancer therapy, analyses of cancer vulnerabilities are increasingly performed at a genome-wide scale and across large panels of in vitro cancer models. We can assemble the joint comprehensive map of all the intracellular genetic dependencies and vulnerabilities of cancer: the Cancer Dependency Map (DepMap).

Studies in human cancer cell lines have accumulated multiple layers of genetic information that can be used to study cancer vulnerabilities. This includes CRISPR KO screens, RNA and drug screens together with gene expression, mutation and copy number variation data. For instance, data from the DepMap Achilles project, based on both CRISPR KO and RNA screens, have been exploited to uncover potential cancer gene dependencies.

To simulate all downstream events following gene suppression, we need a gene regulatory network represented as a directed graph encompassing all genes. We then derived causal relationships from 0 - 1000 breast cancer transcriptomes. Another option is to assign orientations to links in the coexpression network, that represents direct regulatory interactions.

In this work, we sought to develop a computational method that predicts cancer-specific dependencies for clinical samples with breast cancer as a model. We used the results of genome-wide CRISPR KO screening of breast cancer cell lines, the breast cancer regulatory networks described above, and the transcriptomic and mutational profiles of TCGA breast cancer samples. Systematic characterization of common dependencies and experimental validation using patient-derived cells identified novel therapeutic targets in breast cancer. This translational approach lays the foundation for future applications to discovering personalized therapeutic targets from clinical molecular data on various tumor types.

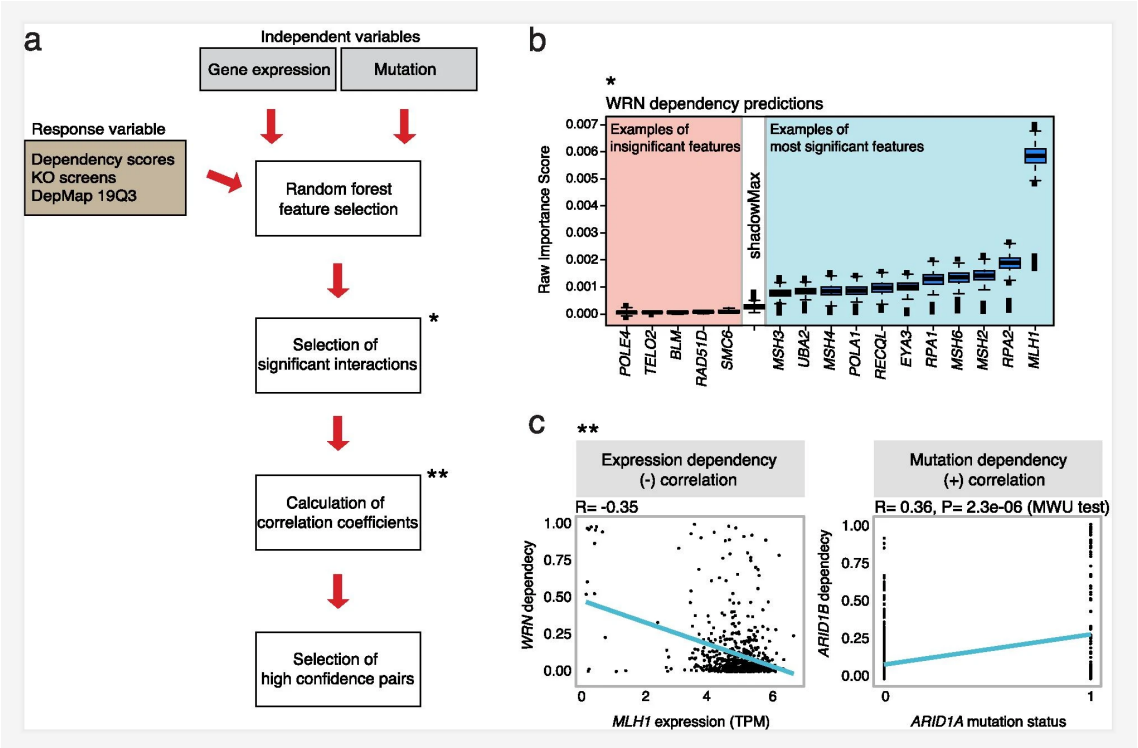

Figure 1: Workflow of machine learning prediction

3 Methods

3.1 Development of CRISPR KO prediction model

Our main work lies in how to clean the vast data. It may take a long time for python to even just read the data, let alone compute. Before we tend the data, we must understand what each data means. It may take us some time on the Biology parts. After cleaning the data, applying each regression may be more accessible.

Transcript protein and matching dependency maps constructed for cancer cell lines are used for training. Tumor and matched normal transcript protein of clinical samples are used as input for prediction. Cancer-specific vulnerability can be identified by comparing the prediction outcomes for tumor and normal samples. The prediction model consists of CRISPR KO screens and machine

learning.

Next step, we will merge two independent CRISPR KO screens of target breast cancer cell lines, each based on a dependency score named CRISPRGeneDependency.csv respectively. The overall goal is to recover the result in the literature listed above. We will use five different models for the data, including random forests, logistic regression, support vector machine, gradient boosting machine, and artificial neural network.

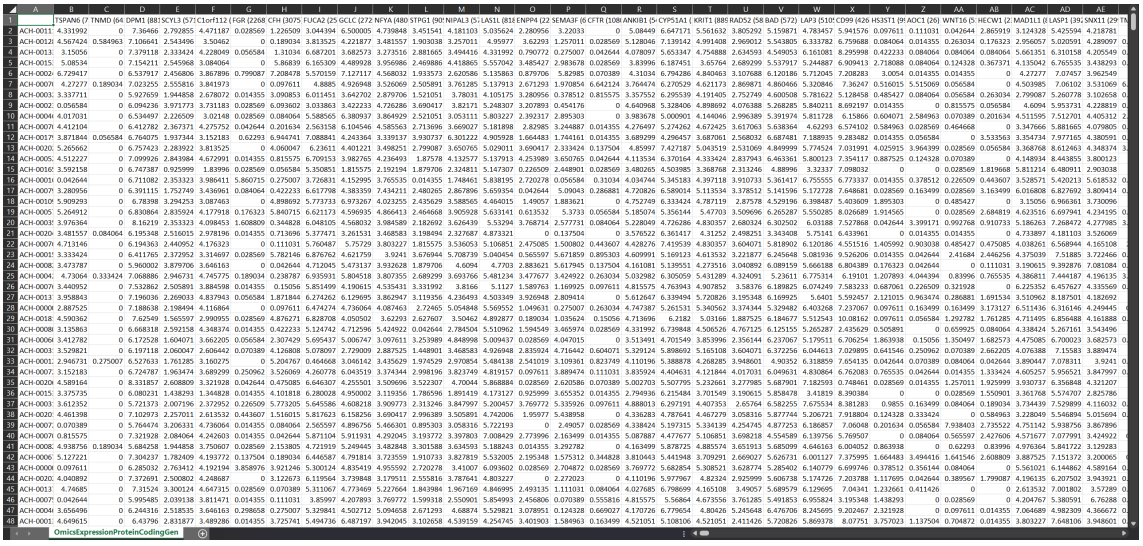

Figure 2: traning gene effects set from OmicsExpressionGenesExpectedCountProfile.csv

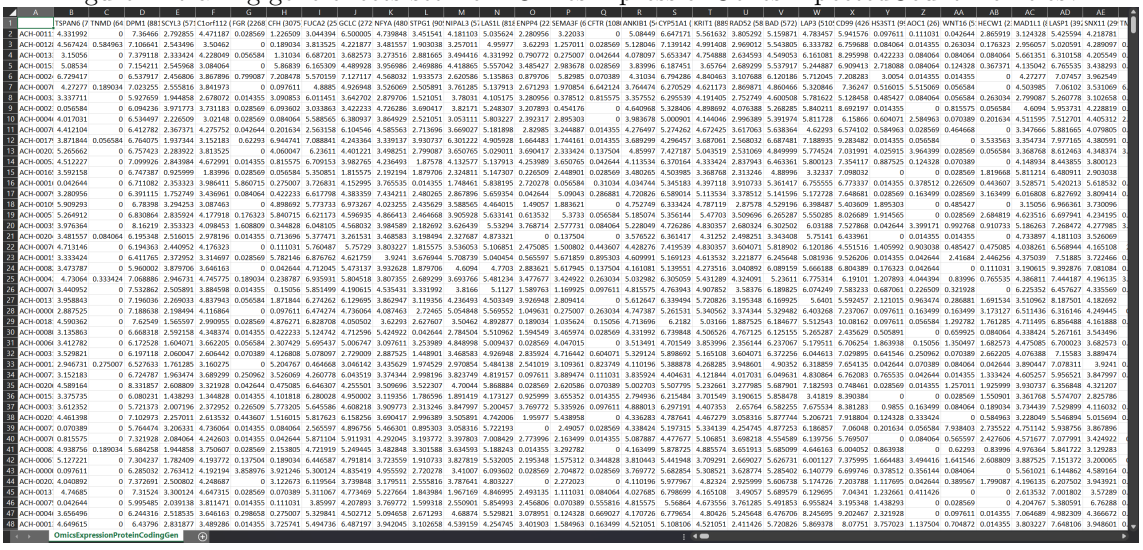

Figure 3: target transcript protein from OmicsExpressionProteinCodingGenesTPMLogp1.csv